Heredity **72** (1994) 219–227

Received 18 June 1993

Estimation of the proportion of diploid males in populations of Hymenoptera

ROBIN E. OWEN* & LAURENCE PACKER†

Department of Chemical and Biological Sciences, Mount Royal College, Calgary, Alberta, Canada T3E 6K6 and Department of Biological Sciences, University of Calgary, Calgary, Alberta, Canada T2N 1N4 and †Department of Biology, York University, North York, Ontario Canada, M3J 1P3

Diploid males occur at low frequencies in natural populations of Hymenoptera as a consequence of the sex-determination system. Routine electrophoretic surveys will often reveal heterozygous diploid males. Maximum likelihood estimates are given for ϕ, the proportion of males in the population that are diploid, when data are available from males only or from both males and females. In the simplest case, using male data only, $\hat{\phi} \approx B_2/2pqT_2$, where p and q are the gene frequencies at the marker locus, B_2 is the number of heterozygous diploid males and T_2 is the total number of males sampled. The variance $V(\hat{\phi}) = \phi[1-2pq\phi-(1-4pq)\phi^2]/2pqT_2$. When both male and female data are available then Φ, the proportion of diploids that are male, can also be estimated. This allows the approximate effective number of sex-determining alleles (assuming a single locus system) to be determined. Maximum likelihood estimates of ϕ have to be obtained numerically when data are available from multiple-allelic or multiple marker loci.

Keywords: diploid males, Hymenoptera, likelihood estimates.

Introduction

Diploid males are expected to occur in natural populations of Hymenoptera as a consequence of the sex-determination mechanism. Heterozygotes at one or more sex-determination loci are female while hemizygotes and homozygotes are male (Crozier, 1971). Although there may be many alleles (9–19; Adams *et al.*, 1977) at these loci, some diploid males will inevitably result each generation. For instance with a single locus system, as found in *Apis mellifera* and *Bracon hebetor*, diploid males will issue from matings between parents having one sex-determining allele in common. These are termed matched matings by Adams *et al.* (1977). As there are only a finite number of alleles present in a population some matched matings will always occur in each generation even with panmixis.

Diploid males themselves are often inviable (Petters & Mettus, 1980) or sterile, moreover they also impose a significant cost on the reproductive success of their parents (Page, 1980; Ross & Fletcher, 1986; Ratnieks, 1990); hence there may be selection for avoidance of inbreeding (Plowright & Pallett, 1979) and for multiple mating by females of social species (Page, 1980).

*Correspondence.

Adams *et al.* (1977) have shown that in an infinite population at equilibrium the frequency of matched matings θ is $2/K$, where K is the effective number of alleles maintained at the sex-determination locus. Therefore, the frequency in the population of diploids that are male, Φ is $(1-s)(\theta/2)$, because only half of the diploid progeny from matched matings are male and where s is the selection coefficient against the diploid males. It is clear that Φ is likely to be small, on the order of 10 per cent or less, in most natural populations. Operationally of course, male diploids are only detected by surveying males, thus the proportion of males in the population that are diploid, ϕ is defined as the number of diploid males divided by the total number of males (diploid + haploid). Therefore ϕ depends not only on the frequency of matched matings but also on the ratio of fertilized to unfertilized eggs, which we will refer to as the primary sex ratio (Fig. 1). Unless the primary sex ratio is highly female (i.e. diploid) biased ϕ, like Φ, will also take values often considerably less than 10 per cent and so it is not at all surprising that although diploid males have been recorded in a number of species of Hymenoptera, usually only a few specimens in each have been found (Crozier, 1971; Kukuk & May, 1990; Packer & Owen, 1990). Surveys of populations using polymorphic gene loci (e.g. allozyme loci) will reveal diploid males, if they

occur, even if the exact system of sex determination is not known (i.e. whether single or multiple loci are involved). Given the considerations discussed above it is clearly desirable to have efficient estimates of ϕ and to have expectations of the effort required to detect diploid males. In the latter context the sample size necessary obviously depends on the parametric value of ϕ and the allele frequencies at the marker locus. In this paper we derive maximum likelihood estimates of the proportion of diploid males for use in situations where data are available from males only or from males and females.

Maximum likelihood estimates

Two alleles: the most general case

The most commonly employed and most useful markers are allozyme loci which generally show co-dominant inheritance; hence heterozygous diploid males can be distinguished but males haploid and homozygous for the same allele have identical electrophoretic phenotypes. Thus at a locus with two alleles, for instance fast (F) and slow (S), there are three phenotypic classes in each sex. Assuming Hardy–Weinberg equilibrium and gene frequencies p and q (Fig. 1) we have:

	Phenotype	F	FS	S	Total
Females:	Number	A_1	B_1	C_1	T_1
	Genotype	FF	FS	SS	
	Frequency	p^2	$2pq$	q^2	
Males	Number	A_2	B_2	C_2	T_2
	Genotypes	$FF+F$	FS	$SS+S$	
	Frequency	$\phi p^2+(1-\phi)p$ $=p(1-q\phi)$	$\phi 2pq$	$\phi q^2+(1-\phi)q$ $=q(1-p\phi)$	

We assume that only one offspring per random mating is included in a sample.

Using male data only. There are two variables to be estimated, ϕ and p. If data are available from males alone then direct estimates of both and their variances are easily obtained. For the marker locus there is clearly only one possible estimate for the allele frequencies, i.e.

$$\hat{p}=(A_2+\tfrac{1}{2}B_2)/T_2 \text{ and } \hat{q}=(\tfrac{1}{2}B_2+C_2)/T_2$$

which is correct regardless of the proportion of diploid males.
Therefore,

$$\hat{\phi}=\frac{B_2}{2\hat{p}\hat{q}T_2} \qquad (1)$$

Note also that $\hat{p}-\hat{p}\hat{q}\hat{\phi}=A_2/T_2$ and $\hat{q}-\hat{p}\hat{q}\hat{\phi}=C_2/T_2$. To find variances we determine the information matrix

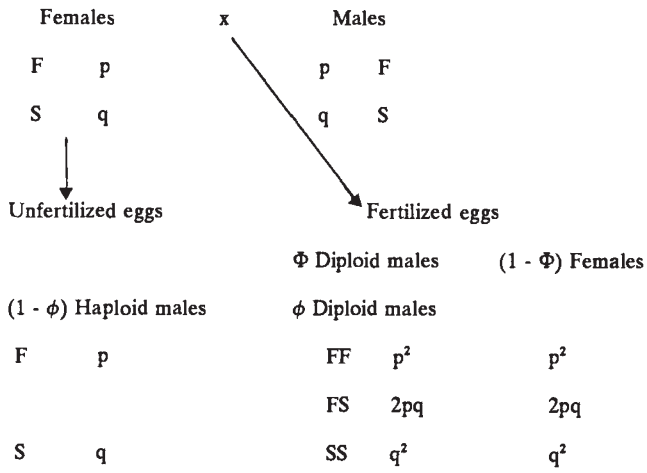

Fig. 1 Origin and frequency of diploid males at a marker locus with alleles F and S at frequencies p and q, respectively, in the parents of the generation under consideration. Diploid males can be viewed in two ways: either as the proportion of diploids that are male (Φ), or as the proportion of males that are diploid (ϕ). ϕ therefore depends on the ratio of fertilized to unfertilized eggs.

$\boldsymbol{I}^{(m)}$. The logarithmic likelihood may be taken as:

$$\begin{aligned}L&=A_2\ln(p-pq\phi)+B_2\ln(pq\phi)+C_2\ln(q-pq\phi)\\&=(A_2+B_2)\ln p+(B_2+C_2)\ln q+A_2\ln(1-q\phi)\\&\quad+B_2\ln\phi+C_2\ln(1-p\phi).\end{aligned} \qquad (2)$$

Taking the second derivatives and inserting the expectations for A_2, B_2 and C_2 gives:

$$\frac{1}{T_2}\boldsymbol{I}^{(m)}_{pp}=\frac{1-2pq\phi-(1-4pq)\phi^2}{pq(1-p\phi)(1-q\phi)}$$

$$\frac{1}{T_2}\boldsymbol{I}^{(m)}_{p\phi}=\boldsymbol{I}^{(m)}_{\phi p}/T_2=-(p-q)(1-\phi)/(1-p\phi)(1-q\phi)$$

$$\frac{1}{T_2}\boldsymbol{I}^{(m)}_{\phi\phi}=\frac{2pq}{\phi}+\frac{pq(1-2pq\phi)}{(1-p\phi)(1-q\phi)}=\frac{(2-\phi)pq}{\phi(1-p\phi)(1-q\phi)}$$

which are the elements of the information matrix $\boldsymbol{I}^{(m)}$. The inverse of $\boldsymbol{I}^{(m)}$ is the variance-covariance matrix, therefore the determinant is required. This can be deduced from the value of $\boldsymbol{I}^{(m)}$ provided $V(\hat{p})$ can be computed otherwise. This is because $V(\hat{p})\cong\boldsymbol{I}^{(m)}_{\phi\phi}/|\boldsymbol{I}^{(m)}|$ so $|\boldsymbol{I}^{(m)}|\cong\boldsymbol{I}^{(m)}_{\phi\phi}/V(\hat{p})$. Now $\hat{p}=(A_2+\frac{1}{2}B_2)/T_2$ which is a linear function of the multinominal frequencies A_2 and B_2, the total number of observations being T_2. Therefore by Fisher's formula (1946):

$$\begin{aligned}T_2V(\hat{p})&=E(A_2/T_2)+\tfrac{1}{4}E(B_2/T_2)-p^2\\&=p-pq\phi+\tfrac{1}{4}(2pq\phi)-p^2=pq(1-\tfrac{1}{2}\phi).\end{aligned}$$

Therefore,

$$V(\hat{p}) = pq(1-\tfrac{1}{2}\phi)/T_2. \quad (3)$$

Hence

$$\frac{|I^{(m)}|}{T_2} \cong \frac{(2-\phi)pq}{\phi(1-p\phi)(1-q\phi)} \cdot \frac{2}{pq(2-\phi)}$$

$$= 2/\phi(1-p\phi)(1-q\phi),$$

which can, of course, be verified by directly calculating $|I^{(m)}|$. The large sample approximation to $V(\hat{\phi})$ is now obtained; we have $T_2 V(\hat{\phi}) = I^{(m)}_{pp} |I^{(m)}|$ giving,

$$V(\hat{\phi})^m = \phi[1-2pq\phi-(1-4pq)\phi^2]/2pqT_2. \quad (4)$$

Eqns 1 and 4 were given previously by Packer & Owen (1990). Note also that Ross & Fletcher (1985) solved the likelihood eqn 2 iteratively to obtain estimates of ϕ and its variance. Our analytical solution obviates the need for this numerical method. It may be noted in passing that substitution of $(1-\alpha)$ for ϕ in eqn 4 gives the large-sample variance estimate of the Bernstein–Wright coefficient α. When $\alpha = 0$, Var $(\hat{\alpha}) = 1/T_2$. Also, eqn 3 is equivalent to $V(\hat{p}) = \frac{1}{2}pq(1+\alpha)$, with $\alpha = 1-\phi$. Goodness of fit cannot be tested because the fit is perfect, the number of parameters equalling the number (two) of independent classes. Given that the frequency of diploid males in most natural populations of Hymenoptera is likely to be low, we can enquire as to the sample size required to detect at least one diploid male. This sample size, which depends on ϕ and the allele frequencies at the marker locus, is given in Table 1. Efficiency of detection increases as the frequencies of the alternative alleles at the marker locus become more equal. It must be noted though, that if only one diploid male is detected then the standard error is the same order of magnitude as the estimate $\hat{\phi}$.

Table 1 The sample size to give an expected number of one heterozygous diploid male as a function of the parametric value of ϕ and the allele frequency p at the marker locus

	p					
ϕ	0.05	0.1	0.2	0.3	0.4	0.5
0.01	1053	556	313	239	209	200
0.02	527	278	157	120	105	100
0.05	211	112	63	48	42	40
0.10	106	56	32	24	21	20
0.20	53	28	16	12	11	10
0.30	36	19	11	8	7	7
0.40	27	14	8	6	6	5
0.50	22	12	7	5	5	4

Using data from males and females. If females have also been scored at the marker locus then this information can be used to improve the estimate of ϕ. In this case we add to the likelihood eqn 2 the item $A_1 \ln p^2 + B_1 \ln(pq) + C_1 \ln q^2$ or $(2A_1 + B_1)\ln p + (B_1 + 2C_1)\ln q$. Therefore for estimation we now have the equations:

$$\frac{\delta L}{\delta p} = \frac{2A_1 + B_1 + A_2 + B_2}{p} - \frac{B_1 + 2C_1 + B_2 + C_2}{q}$$

$$+ \frac{A_2\phi}{1-q\phi} - \frac{C_2\phi}{1-p\phi} = 0, \quad (5)$$

$$\frac{\delta L}{\delta \phi} = -\frac{A_2 q}{1-q\phi} + \frac{B_2}{\phi} - \frac{C_2 p}{1-p\phi} = 0. \quad (6)$$

The number of degrees of freedom of the observations is four, being the number of classes, namely six, less the number of prescribed totals, which is two (T_1 and T_2). The number of parameters under estimation is only two, being p and ϕ. The maximum likelihood equations are unlikely to have an algebraic solution and they will have to be solved numerically. The information matrix is just the original one except that the I_{pp} item is increased by terms arising from the additional item in the logarithmic likelihood. The matrix is therefore

$$I = \begin{pmatrix} \frac{2T_1}{pq} + I^{(m)}_{pp}, & I^{(m)}_{p\phi} \\ I^{(m)}_{\phi p}, & I^{(m)}_{\phi\phi} \end{pmatrix}$$

where $I^{(m)}_{pp}$, etc. indicate the formulae already obtained for the males only. The required large-sample variances and covariances, $V(\hat{p})$, $V(\hat{\phi})$, $Cov(\hat{p}, \hat{\phi})$ are obtained as elements of the variance-covariance matrix which is the inverse of I.

Thus the large sample variance of $\hat{\phi}$ is

$$V(\hat{\phi}) = \frac{T_1}{pqT_2} \frac{\phi(1-p\phi)(1-q\phi)}{(2-\phi)T_1 + T_2} + \frac{T_2}{(2-\phi)T_1 + T_2} V(\phi)^{(m)} \quad (7)$$

where $V(\hat{\phi})^{(m)}$ is given by eqn 4. As only p and ϕ are estimated and there are four independent classes, two degrees of freedom are left over for a chi-squared test of goodness of fit.

When ϕ is small considerable simplification results if we are content with an approximate solution. The maximum likelihood eqns 5 and 6 become approximately:

$$\frac{2A_1 + B_1 + A_2 + B_2}{p} - \frac{B_2 + 2C_1 + B_2 + C_2}{q} = 0$$

giving the estimate

$$\hat{p} = (2A_1 + B_1 + A_2 + B_2)/(2T_1 + T_2) \quad (8)$$

and $B_2/\phi = A_2 q + C_2 p$ with $A_2 \cong T_2\hat{p}$, $C_2 \cong T_2\hat{q}$.

Inserting these latter values into the above equation on the right-hand-side we get approximately

$$B_2/\hat{\phi} = 2T_2\hat{p}\hat{q}$$

i.e. $$\hat{\phi} = B_2/2\hat{p}\hat{q}T_2. \quad (9)$$

This is the same as the previous exact estimation eqn 1 for males only. It is now approximate only, holding as an approximation when ϕ is small.

Correspondingly, the information matrix is represented approximately by:

$$\begin{pmatrix} \frac{2T_1 + T_2}{pq}, & -T_2(p-q) \\ -T_2(p-q), & 2T_2pq/\phi \end{pmatrix}$$

which indicates that

$$V(\hat{p}) \cong pq/(2T_1 + T_2) \quad (10)$$

and

$$V(\hat{\phi}) = \phi/2pqT_2. \quad (11)$$

Frequency of matched matings

In the previous sections we have been concerned with the estimations of ϕ, the proportion of males in the population that are diploid. This is an important parameter and one that is natural to estimate as, by definition, diploid males are only detected when males are surveyed. However, ϕ in fact represents the ratio of diploid males (produced from matched matings) to haploid males which arise independently by parthenogenesis from their mothers. Thus ϕ depends on the primary sex ratio – the ratio of fertilized eggs (giving rise to females and diploid males) to unfertilized eggs (giving rise to haploid males). However, another quantity of fundamental interest to estimate is Φ, the proportion of diploids that are male, because (on the assumption of a single-locus sex-determination system) this will give the frequency of matched matings in the population and hence the effective number of alleles at the sex-determination locus.

If data are available from both males and females then Φ can be estimated. Assume that a total T diploids have been identified, then we have:

	Genotype Number	*FF*	*FS*	*SS*
Females	Observed	A_1	B_1	C_1
	Expected	$(1-\Phi)p^2$	$(1-\Phi)2pqT$	$(1-\Phi)q^2T$
Males	Observed	—	B_2	—
	Expected	$\Phi p^2 T$	$\Phi 2pqT$	$\Phi q^2 T$.

The homozygous diploid males, of course, cannot be distinguished from the corresponding haploid males. Taking the ratio of heterozygous diploid males to heterozygous females gives:

$$\frac{\Phi 2pqT}{(1-\Phi)2pqT} = \frac{B_2}{B_1}.$$

Thus

$$\hat{\Phi} = B_2/(B_1 + B_2) \quad (12)$$

with variance $$V(\hat{\Phi}) = \frac{\hat{\Phi}(1-\hat{\Phi})}{B_1 + B_2}. \quad (13)$$

Although statistically this is a precise estimate in practice it must be interpreted with caution. This is because unbiased sampling of female and male diploids is assumed; it is essential that neither diploid males nor females are proportionally under-represented in the sample.

In laboratory populations of, for instance, parasitoids this is unlikely to be a problem as, in principle, all individuals can be scored. However, prudence should be used when collecting data from natural populations for this purpose. Assuming confidence in the data then from $\hat{\Phi}$ an estimate of the frequency of matched matings and also of the number of sex-determination alleles (for a single-locus system) can be directly obtained. Thus

$$\hat{\theta} = 2\hat{\Phi}/(1-s) \quad (14)$$

and $\hat{K} = 2/\hat{\theta}$.

Because s, the selection coefficient against the diploid males, will in most cases remain unknown these will be minimum estimates.

Application to data

Packer & Owen (1990) found a single heterozygous diploid male at the *Idh* locus in the halictine bee *Augochlorella striata*. Their data are presented below.

	Phenotype:	*F*	*FS*	*S*	
Number	Females	$A_1 = 26$	$B_1 = 11$	$C_1 = 1$	$T_1 = 38$
	Males	$A_2 = 80$	$B_2 = 1$	$C_2 = 25$	$T_2 = 106$

Using male data only. The allele frequencies are $\hat{p} = (80 + 0.5)/106 = 0.759$, $\hat{q} = 0.241$, therefore from eqn 1 $\hat{\phi} = 1/(2 \times 0.759 \times 0.241 \times 106) = 0.0258$, with standard deviation s.d. $= \sqrt{V(\hat{\phi})^{(m)}} = 0.0257$.
Also, from eqn 3 s.d. of $\hat{p} = 0.0415$.

Using data from males and females. The allele frequencies in males and females are not significantly different ($\chi^2_{[1]} = 1.52$, Packer & Owen, 1990). Also the

analysis using the male data only indicates that ϕ is small. Therefore we can use the approximate solution with the combined male and female data to estimate ϕ. Hence from eqns 8 and 10 $\hat{p} = 0.791 \pm 0.032$ and eqns 9 and 11 gives $\hat{\phi} = 0.0286 \pm 0.0286$.

Frequency of matched matings. Almost three times as many males as females were sampled so it is not unreasonable to think that male and female diploids are represented in the proportions in which they occur in the population. Therefore we can make a tentative estimate of Φ. Using eqn 12 $\hat{\Phi} = 1/(11+1) = 0.0833$ (with variance 0.0064). The frequency of matched matings, from eqn 14 and taking $s = 0$, $\hat{\theta} = 0.1667$. The effective number of sex-determination alleles (assuming a single-locus system), $\hat{K} = 2/0.1667 = 12$. It is interesting to note that this latter estimate, although approximate, is consistent with the estimates of the number of sex-determination alleles found in other species of Hymenoptera (Adams *et al.*, 1977; Ross & Fletcher, 1985).

Two alleles: gene frequencies unequal in diploids and haploids

In some cases (e.g. Ross & Fletcher, 1985) allele frequencies at the marker locus are found to be unequal in males and females. The genotypic frequencies of offspring from parents with alleles F and S at frequencies P, Q in males and frequencies p, q in females, respectively, are then:

Phenotype	F	FS	S	Total
Females				
Number	A_1	B_1	C_1	T_1
Genotype	FF	FS	SS	
Frequency	pP	$pQ + qP$	qQ	
Males				
Number	A_2	B_2	C_2	T_2
Genotypes	$FF+F$	FS	$SS+S$	
Frequency	$\phi pP+(1-\phi)p = p(1-Q\phi)$	$\phi(pQ+qP)$	$\phi qQ+(1-\phi)q = q(1-P\phi)$	

The log likelihood is:

$$L = A_1\ln(pP) + B_1\ln(pQ+qP) + C_1\ln(qQ) + A_2\ln[p(1-Q\phi)] + B_2\ln[\phi(pQ+qP)] + C_2\ln[q(1-P\phi)]. \quad (15)$$

The maximum likelihood estimates $\hat{p}$, $\hat{P}$ and $\hat{\phi}$ of p, P and ϕ are the solutions of the simultaneous equations $\delta L/dp = 0$, $\delta L/dP = 0$, $dL/\delta\phi = 0$, i.e.

$$\frac{\delta L}{\delta p} = \frac{A_1+A_2}{p} - \frac{C_1+C_2}{q} - \frac{(B_1+B_2)(P-Q)}{pQ+qP} = 0 \quad (16)$$

$$\frac{\delta L}{\delta P} = \frac{A_1}{P} - \frac{C_1}{Q} + \frac{(B_1+B_2)(p-q)}{pQ+qP} + \frac{A_2\phi}{1-Q\phi} - \frac{C_2\phi}{1-P\phi} = 0 \quad (17)$$

$$\frac{\delta L}{\delta \phi} = -\frac{A_2Q}{1-Q\phi} + \frac{B_2}{\phi} - \frac{C_2P}{1-P\phi} = 0. \quad (18)$$

An algebraic solution does not appear to be available but numerical values for $\hat{p}$, $\hat{P}$ and $\hat{\phi}$ with $\hat{q} = 1-\hat{p}$, $\hat{Q} = 1-\hat{P}$ can be obtained by a variety of numerical processes, available in standard statistical computer packages. Alternatively the EM algorithm (Demster *et al.*, 1977) can be employed.

The large sample variance-covariance matrix of $\hat{p}$, $\hat{P}$ and $\hat{\phi}$ is given by $\boldsymbol{V} = \boldsymbol{I}^{-1}$ the inverse of $\boldsymbol{I}$, the elements of which are given in the Appendix. Therefore variances can be obtained by inserting the estimated values of the parameters into $\boldsymbol{I}$ and inverting.

The number of parameters estimated is three, there are four degrees of freedom, hence one d.f. is left for chi-squared testing of goodness of fit.

Three alleles: male data only

We now consider three alleles at a locus, a situation commonly encountered with allozymes. Assuming Hardy–Weinberg equilibrium, then with three alleles F, M and S at frequencies p, q and r, respectively, we have in males:

Phenotypes	Number	Genotypes	Frequency
F	A	$FF+F$	$p-\phi p(q+r)$
FM	B_1	FM	$2pq\phi$
M	C	$MM+M$	$q-\phi q(p+r)$
FS	B_2	FS	$2pr\phi$
S	D	$SS+S$	$r-\phi r(p+q)$
MS	B_3	MS	$2qr\phi$

Letting $r = 1-p-q$ then the log likelihood equation is:

$$L = (A+B_1+B_2)\ln p + (B_1+C+B_3)\ln q + (B_2+D+B_3)\ln(1-p-q) + (B_1+B_2+B_3)\ln\phi + A\ln[1-\phi(1-p)] + C\ln[1-\phi(1-q)] + D\ln[1-\phi(p+q)], \quad (19)$$

which has partial derivatives,

$$\frac{\delta L}{\delta p} = \frac{(A+B_1+B_2)}{p} - \frac{(B_2+D+B_3)}{(1-p-q)} + \frac{A\phi}{1-\phi(1-p)} - \frac{D\phi}{1-\phi(p+q)}$$

$$\frac{\delta L}{\delta q}=\frac{B_1+C+B_3}{q}-\frac{(B_2+D+B_3)}{(1-p-q)}+\frac{C\phi}{1-\phi(1-q)}$$
$$-\frac{D\phi}{1-\phi(p+q)}$$
$$\frac{\delta L}{\delta \phi}=\frac{(B_1+B_2+B_3)}{\phi}-\frac{A(1-p)}{1-\phi(1-p)}-\frac{C(1-q)}{1-\phi(1-q)}$$
$$-\frac{D(p+q)}{1-\phi(p+q)}.$$

An obvious estimate of ϕ, which is a natural extension of the two-allele case, is

$$\hat{\phi}=\frac{B_1+B_2+B_3}{(2\hat{p}\hat{q}+2\hat{p}\hat{r}+2\hat{q}\hat{r})T}=(B_1+B_2+B_3)/(1-\Sigma p_i^2)T \tag{20}$$

where $\Sigma\hat{p}_i^2=(\hat{p}^2+\hat{q}^2+\hat{r}^2)$,

and $\hat{p}$, $\hat{q}$ and $\hat{r}$ are the gene-counting estimates of p, q and r. However unlike the two-allele case these are not maximum likelihood estimates as it is easy to verify that the likelihood equations are not satisfied. Nevertheless, these are quite good and relatively efficient estimates and provide appropriate starting values for solution of the likelihood equations by numerical techniques, such as the EM method. Once the parameter values have been estimated variances are obtained from the variance-covariance matrix:

$$\begin{pmatrix} V(p) & Cov(p,\phi) & Cov(p,q) \\ Cov(p,\phi) & V(\phi) & Cov(q,\phi) \\ Cov(p,q) & Cov(q,\phi) & V(q) \end{pmatrix}$$
$$=\begin{pmatrix} I_{pp} & I_{p\phi} & I_{pq} \\ I_{p\phi} & I_{\phi\phi} & I_{q\phi} \\ I_{pq} & I_{q\phi} & I_{qq} \end{pmatrix}^{-1}$$

with $V(r)=V(p)+V(q)+2Cov(p,q)$. The elements of the information matrix $\boldsymbol{I}$ are given in the Appendix.

Multiple locus estimates

If individuals are scored at more than one locus then additional information is available with which to estimate ϕ. However, maximum likelihood estimates rapidly become unwieldy when more than two loci are involved.

Consider two loci each with two alleles. Assuming Hardy–Weinberg equilibrium and linkage equilibrium, then the phenotypes, genotypes and their frequencies in males are as given in Table 2. The log likelihood equation is, except for an additive content,

Table 2 Phenotypes, genotypes and expected frequencies of males scored at two loci, each with two alleles at frequencies p, q and u, v respectively

Phenotype	Number	Genotypes	Expected frequency
$F;F$	A_1	$FF+F;\ FF+F$	$p^2u^2\phi+pu(1-\phi)$
$F;FS$	B_1	$FF;\ FS$	$2p^2uv\phi$
$F;S$	A_2	$FF+F;\ SS+S$	$p^2v^2\phi+pv(1-\phi)$
$FS;F$	B_2	$FS;\ FF$	$2pqu^2\phi$
$FS;FS$	B_3	$FS;\ FS$	$4pquv\phi$
$FS;S$	B_4	$FS;\ SS$	$2pqv^2\phi$
$S;F$	C_1	$SS+S;\ FF+F$	$q^2u^2\phi+qu(1-\phi)$
$S;FS$	B_5	$SS;\ FS$	$2q^2uv\phi$
$S;S$	C_2	$SS+S;\ SS+S$	$q^2v^2\phi+qv(1-\phi)$
	T		

$$\begin{aligned}L=&(A_1+A_2+2B_1+B_2+B_3+B_4)\ln p\\&+(C_1+C_2+B_2+B_3+B_4+2B_5)\ln q\\&+(A_1+C_1+B_1+2B_2+B_3+B_5)\ln u\\&+(A_2+C_2+B_1+B_3+2B_4+B_5)\ln v\\&+(B_1+B_2+B_3+B_4+B_5)\ln(\phi)\\&+A_1\ln[1-\phi(1-pu)]+A_2\ln[1-\phi(1-pv)]\\&+C_1\ln[1-\phi(1-qu)]+C_2\ln[1-\phi(1-qv)].\end{aligned} \tag{21}$$

Estimates of the gene frequencies and of ϕ can be obtained by setting the partial derivatives, $\delta L/dp$, $\delta L/\delta u$, $\delta L/\delta\phi$ (given in the Appendix) equal to zero and solving numerically. An approximate estimate of ϕ, which can be used as an initial value is:

$$\hat{\phi}_o=\left(\sum_{i=1}^{5}B\right)/[1-(\hat{p}^2+\hat{q}^2)(\hat{u}^2+\hat{v}^2)]T \tag{22}$$

where $\hat{p}$, $\hat{q}$, $\hat{u}$, $\hat{v}$ are the gene-counting estimates,

i.e. $\hat{p}=[A_1+A_2+B_1+\frac{1}{2}(B_2+B_3+B_4)]/T$, etc.

The variances of the gene counting estimates are

$V(\hat{p})=pq(1-\frac{1}{2}\phi)$, $V(\hat{u})=uv(1-\frac{1}{2}\phi)$.

The covariance $Cov(\hat{p},\hat{u})$ is zero. Other variances are then obtained by inserting the final estimates into the information matrix, the elements of which are given in the Appendix.

Clearly it is hardly worthwhile deriving maximum likelihood estimates for more than two loci. Single locus estimates, with their variances, therefore should be made for each locus separately. The drawback here is that obviously at some, if not many, of the loci no hetero-

zygous diploid males will occur and so information from these loci cannot be used.

An alternative approach, suggested by Kukuk & May (1990) is to combine data over all loci and estimate ϕ using an equation equivalent to:

$$\hat{\phi}_o = \frac{(\Sigma B)}{[1 - \Pi(\Sigma p_i^2)]T}. \quad (23)$$

Eqn 23, equivalent to Kukuk & May eqn 1, reduces to eqn 22 for the case of two loci each with two alleles and to eqn 1 for the case of a single locus. But, as pointed out above, eqn 23, except when applied to a single locus, is not a maximum likelihood estimate. However, when ϕ is small eqn 23 will give an approximate estimate which is in fact quite good and its variance will only slightly exceed that of the M.L. estimate. However the use of the binomial variance ($V(\hat{\phi}_o) = \phi/(1 - \Pi)T$), as performed by Kukuk & May (1990), is erroneous and will give a considerable underestimate of the true variance.

It should be stressed that eqn 23 in its general form covers all cases and provides a good estimate of ϕ, or at least a good starting value.

Recommendations

Diploid males are expected to occur at low frequencies in most natural populations of Hymenoptera. Routine electrophoretic surveys will often detect heterozygous diploid males. Even if only one is found it nevertheless reveals valuable comparative information about the genetic structure of the population or species under consideration. Because data from males will be available initial estimates of allele frequencies, ϕ and their variances can be made using eqns 1, 3 and 4. For most purposes this will be sufficient. However, if there is interest in investigating further, then data from females can also be used. If allele frequencies at the marker locus do not differ between males and females then exact estimates of p and ϕ can be obtained by numerically iterating eqns 5 and 6 with the variance of ϕ given by eqn 7. Alternatively, if ϕ is small approximations are given directly by eqns 8–11. If allele frequencies do differ between males and females then numerical iteration of eqns 16–18 is required to obtain $\hat{p}$, $\hat{P}$ and $\hat{\phi}$.

If unbiased sampling of males and females has been carried out then the proportion of diploids that are male (Φ) can be estimated using eqn 12 and from this the frequency of matched matings (eqn 13) and the effective number of sex-determination alleles can be determined. Maximum likelihood estimates of ϕ have to be obtained numerically for the case of multiple (3) alleles and multiple loci, however good approximations are given by eqns 20 and 23, respectively.

Diploid males represent accidental male production – eggs are fertilized as if to give rise to a diploid female individual but because of homozygosity at a sex-determining locus a male is produced. Diploid male production in a social insect reproductive brood really represents an attempt at female production and diploid males should be counted as investment in the female sex and not in the male.

The fact that over 10 per cent of the males in a population may be diploid (Kukuk & May, 1990) suggests that substantial biases can be made in empirical sex ratio studies by the inclusion of diploid males as males rather than females.

Therefore attempts should be made to estimate the proportion of diploid males as accurately as possible.

Acknowledgements

We thank Dr A. R. G. Owen for his comments on the manuscript and the Natural Sciences and Engineering Research Council of Canada for research funds.

References

ADAMS, J., ROTHMAN, E. D., KERR, W. E. AND PAULINO, Z. L. 1977. Estimation of the number of sex alleles and queen matings from diploid male frequencies in a population of *Apis mellifera*. *Genetics*, **86**, 583–596.

CROZIER, R. H. 1971. Heterozygosity and sex determination in haplo-diploidy. *Am. Nat.*, **105**, 399–412.

DEMPSTER, A. P., LAIRD, N. M. AND RUBIN, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc.*, **39**, 1–22.

FISHER, R. A. 1946. A system of scoring linkage data, with special reference to the pied factors in mice. *Am. Nat.*, **80**, 568–578.

KUKUK, P. AND MAY, B. 1990. Diploid males in a primitively eusocial bee, *Lasioglossum* (*Dialictus*) *zephyrum* (Hymenoptera: Halictidae). *Evolution*, **44**, 1522–1528.

PACKER, L. AND OWEN, R. E. 1990. Allozyme variation, linkage disequilibrium and diploid male production in a primitively social bee *Augochlorella striata* (Hymenoptera; Halictidae). *Heredity*, **65**, 241–248.

PAGE, R. E. 1980. The evolution of multiple mating behaviour by honey bee queens (*Apis mellifera* L.). *Genetics*, **96**, 263–273.

PETTERS, R. M. AND METTUS, R. V. 1980. Decreased diploid male viability in the parasitic wasp, *Bracon hebetor*. *J. Hered.*, **71**, 353–356.

PLOWRIGHT, R. C. AND PALLETT, M. J. 1979. Worker-male conflict and inbreeding in bumble bees (Hymenoptera: Apidae). *Can. Ent.*, **111**, 289–294.

RATNIEKS, F. L. W. 1990. The evolution of polyandry by queens in social Hymenoptera: the significance of the timing of removal of diploid males. *Behav. Ecol. Sociobiol.*, **26**, 343–348.

ROSS, K. G. AND FLETCHER, D. J. C. 1985. Genetic origin of male diploidy in the fire ant, *Solenopsis invicta* (Hymenoptera: Formicidae) and its evolutionary significance. *Evolution*, **39**, 888–903.

ROSS, K. G. AND FLETCHER, D. J. C. 1986. Diploid male production – a significant colony mortality factor in the fire ant, *Solenopsis invicta* (Hymenoptera: Formicidae). *Behav. Ecol. Sociobiol.*, **19**, 283–291.

Appendix

The elements of the information matrices are the expected values of the second derivatives of the log likelihood equation. The expected values are obtained by replacing the observed numbers with the corresponding expected numbers.

The inverse of an information matrix $(\boldsymbol{I})^{-1}$ is the variance-covariance matrix.

Two alleles: different allele frequencies in males and females

The elements of $\boldsymbol{I}$ are:

$$\boldsymbol{I}_{pp} = T_1\left(\frac{pQ^2+qP^2}{pq(pQ+qP)}\right) + T_2\left(\frac{1}{pq} - \frac{\phi PQ}{pq(pQ+qP)}\right)$$

$$\boldsymbol{I}_{Pp} = T_1\left(\frac{Pq^2+Qp^2}{PQ(pQ+qP)}\right) + T_2\left(\frac{\phi(p-q)^2}{pQ+qP} + \frac{\phi^2 p}{1-Q\phi} + \frac{\phi^2 q}{1-P\phi}\right)$$

$$\boldsymbol{I}_{\phi\phi} = T_2\left(\frac{pQ+qP-PQ\phi}{\phi(1-\phi+PQ\phi^2)}\right)$$

$$\boldsymbol{I}_{pP} = \boldsymbol{I}_{Pp} = T_1\left(\frac{1}{pQ+qP}\right) + T_2\left(\frac{\phi}{pQ+qP}\right)$$

$$\boldsymbol{I}_{p\phi} = \boldsymbol{I}_{\phi p} = 0$$

$$\boldsymbol{I}_{P\phi} = \boldsymbol{I}_{\phi P} = T_2\left(\frac{q}{1-P\phi} - \frac{p}{1-Q\phi}\right).$$

Three alleles; male data only

The elements of $\boldsymbol{I}$ are:

$$T^{-1}\boldsymbol{I}_{pp} = \frac{1-q}{pr} + \phi\left(\frac{1}{p}+\frac{1}{r}-2\right) + \frac{\phi^2 p}{1-\phi(1-p)} + \frac{\phi^2 r}{1-\phi(1-r)}$$

$$T^{-1}\boldsymbol{I}_{qq} = \frac{1-p}{qr} + \phi\left(\frac{1}{q}+\frac{1}{r}-2\right) + \frac{\phi^2 q}{1-\phi(1-q)} + \frac{\phi^2 r}{1-\phi(1-r)}$$

$$T^{-1}\boldsymbol{I}_{\phi\phi} = \frac{2(pq+qr+pr)}{\phi} + \frac{p(1-p)^2}{1-\phi(1-p)} + \frac{q(1-q)^2}{1-\phi(1-q)} + \frac{r(1-r)^2}{1-\phi(1-r)}$$

$$T^{-1}\boldsymbol{I}_{pq} = \frac{1}{r} + \frac{\phi(1-r)}{r} + \frac{\phi^2 r}{1-\phi(1-r)}$$

$$T^{-1}\mathbf{I}_{p\phi} = \frac{r}{1-\phi(1-r)} - \frac{p}{1-\phi(1-r)}$$

$$T^{-1}\boldsymbol{I}_{q\phi} = \frac{q}{1-\phi(1-q)} - \frac{r}{1-\phi(1-r)}.$$

Two loci; male data only

The partial derivatives of the log likelihood eqn 21 are:

$$\delta L/\delta p = (A_1+A_2+2B_1+B_2+B_3+B_4)/p - (C_1+C_2+2B_5+B_2+B_3+B_4)q + A_2\phi u/k + A_2\phi v/l - C_1\phi u/m - C_2\phi v/n$$

$$\delta L/\delta u = (A_1+C_1+2B_2+B_1+B_2+B_3+B_5)/u - (A_2+C_2+2B_4+B_1+B_3+B_5)/v + A_1\phi p/k - A_2\phi p/l + C_1\phi q/m - C_2\phi q/n$$

$$\delta L/\delta\phi = (B_1+B_2+B_3+B_4+B_5)/\phi - A_1(1-pu)/k - A_2(1-pv)/1 - C_1(1-qu)/m - C_2(1-qv)/n$$

where

$k = 1-\phi(1-pu)$

$l = 1-\phi(1-pv)$

$m = 1-\phi(1-qu)$

$n = 1-\phi(1-qv).$

The elements are $\boldsymbol{I}$ are:

$$T^{-1}\boldsymbol{I}_{pp} = 1/pq + \phi(p^2+q^2)/pq + 4\phi uv + \phi^2 pu^3/k + \phi^2 pv^3/l + \phi^2 qu^3/m + \phi^2 qv^3/n$$

$$T^{-1}\boldsymbol{I}_{uu} = 1/uv + \phi^2(u^2+v^2)/uv + 4\phi pq + \phi p^3 u/k + \phi^2 p^3 v/l + \phi^2 q^3 u/m + \phi^2 q^3 v/n$$

$$T^{-1}\boldsymbol{I}_{\phi\phi} = [1-(p^2+q^2)(u^2+v^2)]/\phi + pu(1-pu)^2/k + pv(1-pr)^2/l + qu(1-qu)^2/k + qv(1-qv)^2/l$$

$$T^{-1}\boldsymbol{I}_{pu} = T^{-1}\boldsymbol{I}_{up} = -\phi(1-\phi)pu/k + \phi(1-\phi)pv/l + \phi(1-\phi)qu/m - \phi(1-\phi)qv/n$$

$$T^{-1}\boldsymbol{I}_{p\phi} = T^{-1}\boldsymbol{I}_{\phi p} = -pu^2/k - pv^2/l + qu^2/m + qv^2/n$$

$$T^{-1}\boldsymbol{I}_{u\phi} = T^{-1}\boldsymbol{I}_{\phi u} = -p^2 u/k + p^2 v/l - q^2 u/m + q^2 v/n.$$